doi:10.1093/nar/gkn317

BioLit: integrating biological literature with databases

J. Lynn Fink*, Sergey Kushch, Parker R. Williams and Philip E. Bourne

Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, CA, 92093-0444, USA

Received January 30, 2008; Revised April 22, 2008; Accepted May 7, 2008

ABSTRACT

BioLit is a web server which provides metadata describing the semantic content of all open access, peer-reviewed articles which describe research from the major life sciences literature archive, PubMed Central. Specifically, these metadata include database identifiers and ontology terms found within the full text of the article. BioLit delivers these metadata in the form of XML-based article files and as a custom web-based article viewer that provides context-specific functionality to the metadata. This resource aims to integrate the traditional scientific publication directly into existing biological databases, thus obviating the need for a user to search in multiple locations for information relating to a specific item of interest, for example published experimental results associated with a particular biological database entry. As an example of a possible use of BioLit, we also present an instance of the Protein Data Bank fully integrated with BioLit data. We expect that the community of life scientists in general will be the primary end-users of the web-based viewer, while biocurators will make use of the metadata-containing XML files and the BioLit database of article data. BioLit is available at http://biolit.ucsd.edu.

INTRODUCTION

Prior to the advent and subsequent popularity of the internet, biological databases and scientific publications were necessarily separate entities. However, now that scientists rely on cyberinfrastructure in the course of their daily research, it is startling that databases and publications are still so distinct from each other. While databases have matured and grown in complexity in the digital age, publications have fared poorly; the traditional concept of an article as a static, independent unit of information persists, albeit in an electronic version. There are myriad ways in which the content of an electronic version of an article can be used in a computational manner and the technology and infrastructure to make this happen already exists; this is demonstrated by, among other things, the existence of biological databases and related data mining tools. Indeed, one hallmark of a successful biological database is integration via data mining with other related databases or resources. Yet integration with literature, which is unquestionably the primary medium through which scientists communicate their research, is conspicuously lacking.

Significant progress in taking advantage of article content is finally being made, perhaps most notably as a result of the open access literature movement pioneered by, among others, the Public Library of Science (PLoS) (1). Articles that are published under an open access license are available for download *in toto* immediately upon publication and can be distributed freely providing that the original citation is maintained. The deposition of open access life sciences articles has been centralized with the creation of PubMed Central (2) and it is especially important to note that PubMed Central stores many of these articles in a standardized and machine-readable format, which conforms to the National Library of Medicine (NLM) DTD (http://dtd.nlm.nih.gov/). All PubMed Central articles deposited as a result of this movement are thus freely available as XML files, which contain the full-text of the article and some semantic mark-up of the content.

Though a major advantage to both authors and readers, a large segment of the life sciences community remains unaware of the opportunities that are now available due to the existence of PubMed Central as a digital repository, or even of the existence of open access literature altogether. Hopefully, this will change with the recent NIH directive, which mandates the deposition of all findings of NIH-funded research in PubMed Central (http://publicaccess.nih.gov/) and with novel open access-driven applications such as PubNet (3) and SciVee (4). Admittedly, challenges do present themselves when considering the use of open access literature beyond that of a reader or distributor—they have been described elsewhere (4–7)—but opportunities abound and the tenor of current scientific policy and inquiry suggests that the open access corpus will continue to grow significantly (8–10).

*To whom correspondence should be addressed. Tel: +1 858 822 1897; Fax: + 1 858 822 3610; Email: jlfink@ucsd.edu

BioLit aims to take advantage of the opportunities offered by open access literature by making it possible to include full text or excerpts of these articles directly within existing biological databases and to add newly generated metadata to the articles in order to increase their informative value, tasks which are both made possible by having free access to the full text of the article. These tasks are accomplished by mining the full article text for terms of interest, indexing the terms found, and then including these terms in both machine-readable and human-readable forms that are an enhancement of the original article XML file. The machine-readable form can be incorporated directly into a database or resource. We have prototyped this particular aspect of the effort using a clone of the RCSB Protein Data Bank (PDB) (11), a database of macromolecular structures, and have consequently focused on finding PDB identifiers and Gene Ontology (12) terms as these are both major resources in the structural biology community and the life sciences community as a whole.

Several groups have made significant contributions in applying ontologies to both open- and closed-access biomedical literature. For example, the BioCreAtIvE initiative has spawned a number of tools with the aim of annotating genes and proteins in articles using ontologies and other vocabularies (13). GoPubMed (14) and the more recent SEGOPubmed (15) both aim to add Gene Ontology-based semantic data to PubMed abstracts in order to improve literature searches. The Textpresso team has also focused on improving the classification and searchability of articles by inferring semantic relationships in articles using customized ontologies (16,17). Unlike GoPubMed and SEGOPubmed, they use the full text of the article. Another excellent resource is the National Center for Biomedical Ontology (NCBO) whose overarching aim is to render biomedical information into machine-readable data (18). One of their major contributions is the unification and integration of biomedical ontologies into a single resource, the OBO Foundry, which makes projects such as these and BioLit feasible. The NCBO consortium is also developing a number of web-based tools, in order to make the machine-readable biomedical information available and useful to the community. Finally, The Conceptual Open Hypermedia Service (COHSE), perhaps the most broadly applicable semantic mark-up project, runs as a portlet and adds hyperlinks to existing web pages by allowing a user to select an ontology with terms that can be matched to the page to create links to other pages (19).

Several of the aforementioned tools allow manual curation of the marked-up result in order to increase the accuracy of the mark-up, and hence, the literature search results. This is the ideal approach towards semantic mark-up since the author, or at least an expert human, is providing input. However, the uptake of such an approach in the community has been fairly unsuccessful owing to time limitations on the part of the putative curators and, presumably, lack of demonstrable value to justify the time and effort.

BioLit shares many attributes with these projects but is unique in that it offers a searchable web database of the *full text* of *all PubMed Central research articles* with automated mark-up from *multiple* biomedical ontologies and identifiers from *multiple* biological databases. *No downloading* of software is required and all searchable terms belong to established biomedical resources. The focus of this resource is specific to life sciences literature but broad within that realm. Machine-readable files are freely available and distributable via automated means.

BIOLIT DATABASE

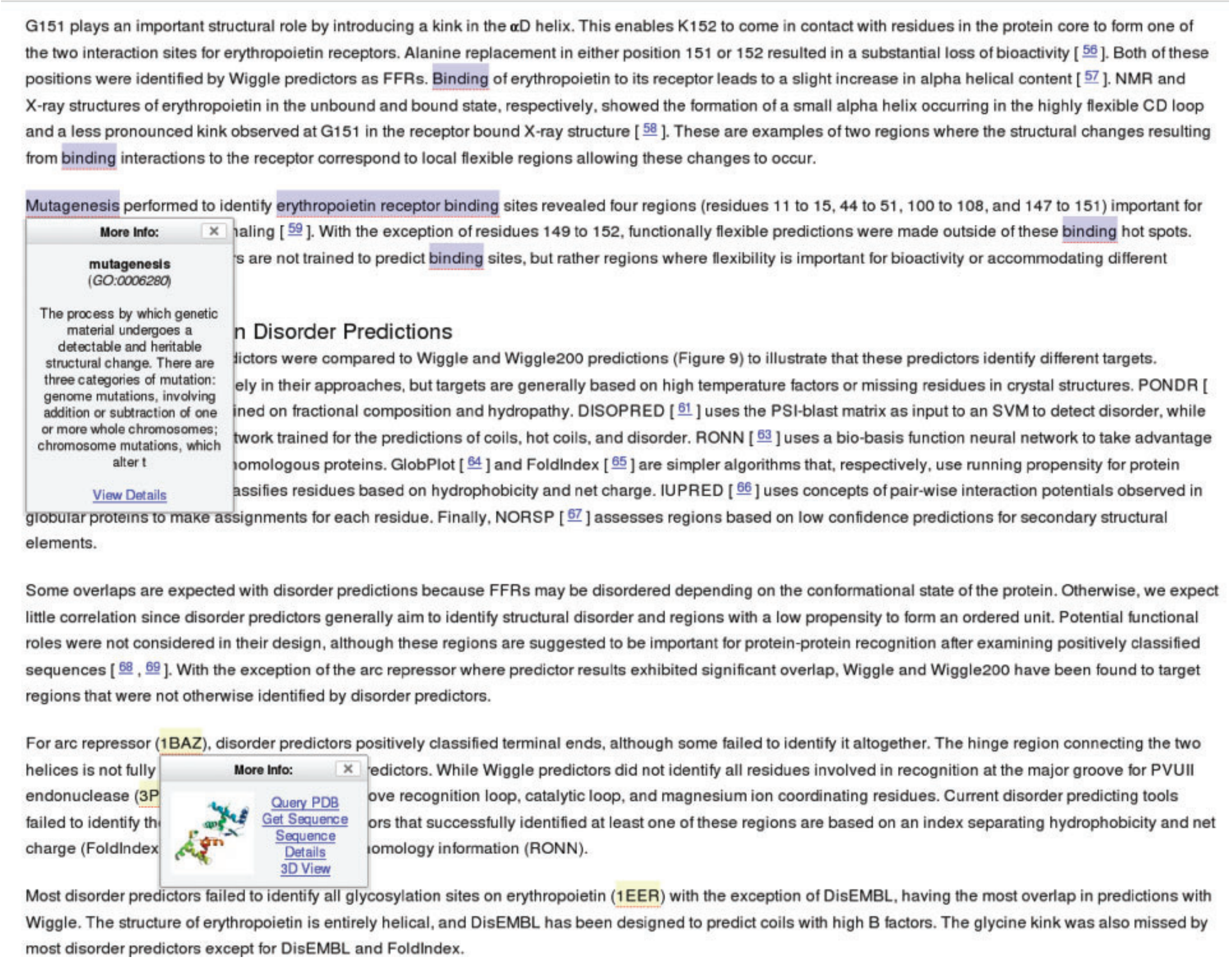
The database supporting the BioLit web server contains data from a subset of the PubMed Central holdings. Article archives, including full-text XML files and figures, were retrieved via the PubMed Central FTP site (ftp://ftp.ncbi.nlm.nih.gov/pub/pmc). The articles were first filtered to remove articles that were not labeled as ‘review-article’, ‘case-report’, ‘research-article’, ‘brief-report’, ‘retraction’ and ‘article-commentary’, and documents which contained only scanned content since these are generally advertisements. The remaining articles were stored in a local MySQL database both in the original and parsed formats. When possible, data included in the MEDLINE record for an article were retrieved from PubMed using NCBI’s E-Utilities (2). The full text of each article was then parsed to find PDB IDs and Gene ontology terms. These metadata were also stored in the local database. Currently, 51 667 articles exist in the database. The database is updated weekly.

BIOLIT ARTICLE VIEWER

BioLit uses a customized viewer for the articles in order to display the original text inline with the added metadata. This viewer uses a custom XML stylesheet based on the NLM DTD. The full text of the paper can be displayed in a tabbed format or the traditional linear article format and the metadata are highlighted with color and a menu of options when the user clicks on a term. In the case of PDB IDs, an image of the 3D structure identified by the PDB ID is displayed along with links to the macromolecular sequence, the PDB record and other related features. Clicking on an ontology term shows a definition of the term and related links. We also generate statistics describing ontology term usage across all articles and these terms can be used for searching or finding related articles. Figure 1 shows an excerpt of BioLit-enhanced text.

In order to use the BioLit article viewer, the user enters a search term on the BioLit homepage. This term can be an author name, title, keyword, etc. The web server then returns a list of matching articles; clicking on an article invokes the customized viewer. Alternatively, a user could bypass the web-based viewer and retrieve BioLit XML files using a script by appending a PubMed Central ID (PMCID), PubMed ID (PMID) or article digital object identifier (DOI) to the following URL template: http://biolit.ucsd.edu/biolit/getbiolitxml?SOURCE = ID where, SOURCE is ‘pmcid’, ‘pmid’ or ‘doi’ and ID is the corresponding identifier for that resource.

The machine-readable version of an article is available as an XML file through the viewer and via a POST



G151 plays an important structural role by introducing a kink in the αD helix. This enables K152 to come in contact with residues in the protein core to form one of the two interaction sites for erythropoietin receptors. Alanine replacement in either position 151 or 152 resulted in a substantial loss of bioactivity [56]. Both of these positions were identified by Wiggle predictors as FFRs. Binding of erythropoietin to its receptor leads to a slight increase in alpha helical content [57]. NMR and X-ray structures of erythropoietin in the unbound and bound state, respectively, showed the formation of a small alpha helix occurring in the highly flexible CD loop and a less pronounced kink observed at G151 in the receptor bound X-ray structure [58]. These are examples of two regions where the structural changes resulting from binding interactions to the receptor correspond to local flexible regions allowing these changes to occur.

Mutagenesis performed to identify erythropoietin receptor binding sites revealed four regions (residues 11 to 15, 44 to 51, 100 to 108, and 147 to 151) important for
naling [59]. With the exception of residues 149 to 152, functionally flexible predictions were made outside of these binding hot spots.
s are not trained to predict binding sites, but rather regions where flexibility is important for bioactivity or accommodating different

More Info:

mutagenesis
(GO:0006280)

The process by which genetic material undergoes a detectable and heritable structural change. There are three categories of mutation: genome mutations, involving addition or subtraction of one or more whole chromosomes; chromosome mutations, which alter t

View Details

n Disorder Predictions

dictors were compared to Wiggle and Wiggle200 predictions (Figure 9) to illustrate that these predictors identify different targets.
ely in their approaches, but targets are generally based on high temperature factors or missing residues in crystal structures. PONDR [
ined on fractional composition and hydropathy. DISOPRED [61] uses the PSI-blast matrix as input to an SVM to detect disorder, while
twork trained for the predictions of coils, hot coils, and disorder. RONN [63] uses a bio-basis function neural network to take advantage
omologous proteins. GlobPlot [64] and FoldIndex [65] are simpler algorithms that, respectively, use running propensity for protein
assifies residues based on hydrophobicity and net charge. IUPRED [66] uses concepts of pair-wise interaction potentials observed in
globular proteins to make assignments for each residue. Finally, NORSP [67] assesses regions based on low confidence predictions for secondary structural elements.

Some overlaps are expected with disorder predictions because FFRs may be disordered depending on the conformational state of the protein. Otherwise, we expect little correlation since disorder predictors generally aim to identify structural disorder and regions with a low propensity to form an ordered unit. Potential functional roles were not considered in their design, although these regions are suggested to be important for protein-protein recognition after examining positively classified sequences [68 , 69]. With the exception of the arc repressor where predictor results exhibited significant overlap, Wiggle and Wiggle200 have been found to target regions that were not otherwise identified by disorder predictors.

For arc repressor (1BAZ), disorder predictors positively classified terminal ends, although some failed to identify it altogether. The hinge region connecting the two
helices is not fully
edictors. While Wiggle predictors did not identify all residues involved in recognition at the major groove for PVUII
endonuclease (3P
ove recognition loop, catalytic loop, and magnesium ion coordinating residues. Current disorder predicting tools
failed to identify th
ors that successfully identified at least one of these regions are based on an index separating hydrophobicity and net
charge (FoldIndex
omology information (RONN).

Most disorder predictors failed to identify all glycosylation sites on erythropoietin (1EER) with the exception of DisEMBL, having the most overlap in predictions with Wiggle. The structure of erythropoietin is entirely helical, and DisEMBL has been designed to predict coils with high B factors. The glycine kink was also missed by most disorder predictors except for DisEMBL and FoldIndex.

Figure 1. A screenshot of the BioLit article viewer. This customized viewer not only displays the full text of an article as it was originally published but also displays integrated metadata from the BioLit database. PDB IDs can be seen highlighted in yellow and Gene Ontology terms are highlighted in lavender. Clicking on a highlighted terms brings up information, a list of options or links associated with each term. Article shown is in *PLoS Comput. Biol.*, 2006, **2**, e90.

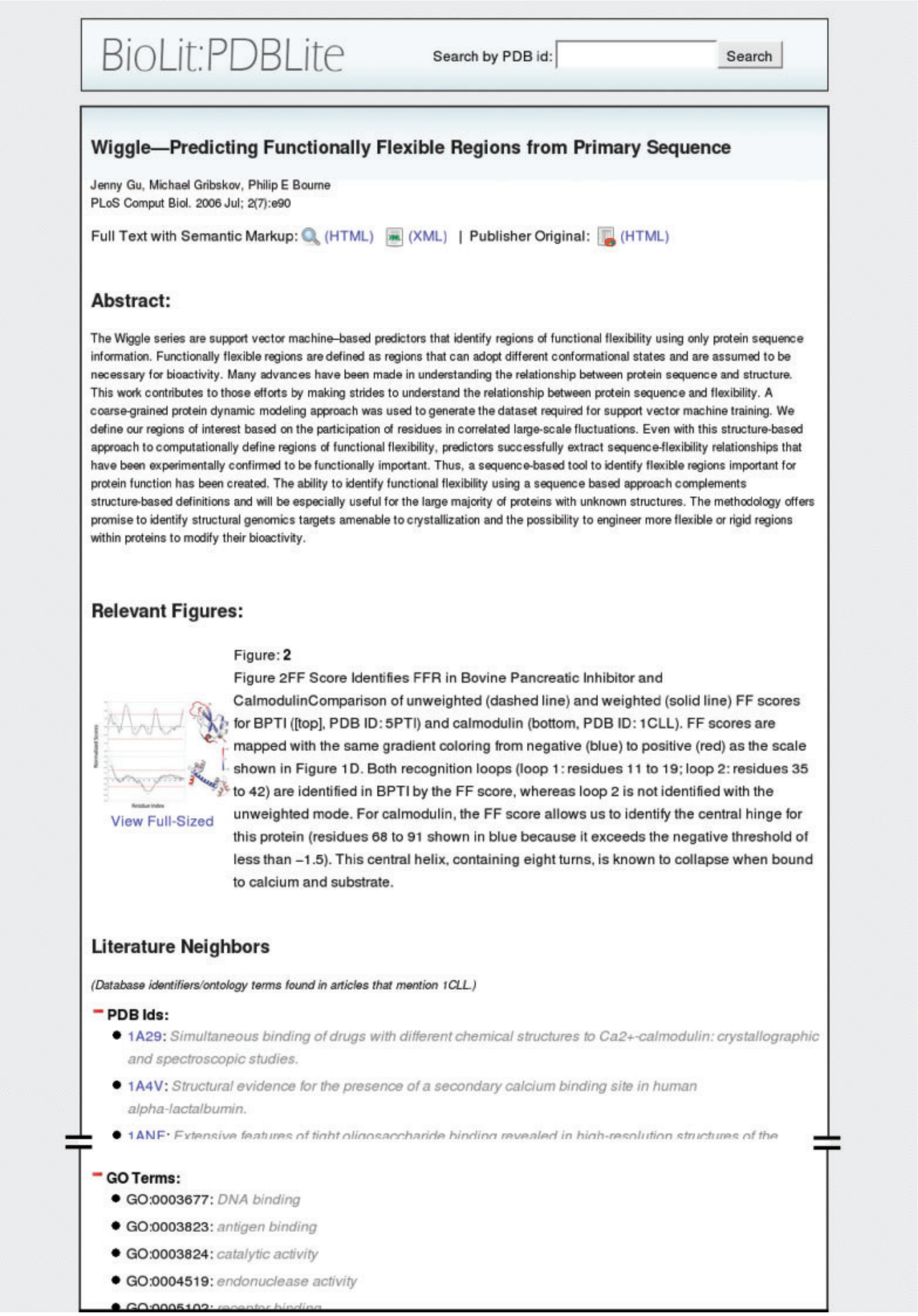
HTTP query. The XML markup enrichment performed by BioLit consists of inserting metadata into the original XML file. The <*biolit*> tag appears as the top-level child node of <*article*>, a tag in the NLM namespace. The additional metadata includes markup revision data (<*markup-version*> and <*markup-date*>) as well as derived identification content <*ident*>, which specifies the DOI and PubMed Central ID of the article. Throughout the text of the article, <*biolit-term*> marks semantic content that has been discovered during the data mining process where attributes type, id identify the content. The complete BioLit XML specification can be found at http://biolit.ucsd.edu.

BIOLIT-INTEGRATED PROTEIN DATA BANK CLONE

One example of how the BioLit data might be used in an existing biological web resource is illustrated by the BioLit PDB clone, a stand-alone clone of the Protein Data Bank which runs independently from the PDB and has direct access to the BioLit database (http://biolit.ucsd.edu/pdb). The clone maintains much of the original functionality, including access to the latest release of the PDB database, while providing access to features unique to BioLit. The PDB structure explorer page now shows related articles as identified by BioLit. A simple browser has been integrated to display data excerpts from the article as well as structure IDs and ontology terms identified in the article. Thus, when an entry in the PDB is displayed, all articles in the BioLit database that mention that entry are included as citations in the PDB record. In contrast, the PDB currently only displays the citation associated with the entry in which the solved structure is reported. Figure 2 shows an entry from the BioLit PDB clone which has citations in the BioLit database. Because we find all mentions of PDB IDs in an article, we can also report which PDB IDs are mentioned in the same article. This can direct the user to other PDB entries which might be of interest. The user can also access the BioLit article viewer via the PDB record.

ONTOLOGY PARSING

All ontology terms are parsed from an OBO file (http://obofoundry.org/ro/ro.obo) and loaded into a tree



Wiggle—Predicting Functionally Flexible Regions from Primary Sequence

Jenny Gu, Michael Gribskov, Philip E Bourne
PLoS Comput Biol. 2006 Jul; 2(7):e90

Full Text with Semantic Markup: (HTML) (XML) | Publisher Original: (HTML)

Abstract:

The Wiggle series are support vector machine–based predictors that identify regions of functional flexibility using only protein sequence information. Functionally flexible regions are defined as regions that can adopt different conformational states and are assumed to be necessary for bioactivity. Many advances have been made in understanding the relationship between protein sequence and structure. This work contributes to those efforts by making strides to understand the relationship between protein sequence and flexibility. A coarse-grained protein dynamic modeling approach was used to generate the dataset required for support vector machine training. We define our regions of interest based on the participation of residues in correlated large-scale fluctuations. Even with this structure-based approach to computationally define regions of functional flexibility, predictors successfully extract sequence-flexibility relationships that have been experimentally confirmed to be functionally important. Thus, a sequence-based tool to identify flexible regions important for protein function has been created. The ability to identify functional flexibility using a sequence based approach complements structure-based definitions and will be especially useful for the large majority of proteins with unknown structures. The methodology offers promise to identify structural genomics targets amenable to crystallization and the possibility to engineer more flexible or rigid regions within proteins to modify their bioactivity.

Relevant Figures:

Figure: **2**
Figure 2FF Score Identifies FFR in Bovine Pancreatic Inhibitor and CalmodulinComparison of unweighted (dashed line) and weighted (solid line) FF scores for BPTI ([top], PDB ID: 5PTI) and calmodulin (bottom, PDB ID: 1CLL). FF scores are mapped with the same gradient coloring from negative (blue) to positive (red) as the scale shown in Figure 1D. Both recognition loops (loop 1: residues 11 to 19; loop 2: residues 35 to 42) are identified in BPTI by the FF score, whereas loop 2 is not identified with the unweighted mode. For calmodulin, the FF score allows us to identify the central hinge for this protein (residues 68 to 91 shown in blue because it exceeds the negative threshold of less than −1.5). This central helix, containing eight turns, is known to collapse when bound to calcium and substrate.

View Full-Sized

Literature Neighbors

(Database identifiers/ontology terms found in articles that mention 1CLL.)

− **PDB Ids:**

- 1A29: *Simultaneous binding of drugs with different chemical structures to Ca2+-calmodulin: crystallographic and spectroscopic studies.*
- 1A4V: *Structural evidence for the presence of a secondary calcium binding site in human alpha-lactalbumin.*
- 1ANE: *Extensive features of tight oligosaccharide binding revealed in high-resolution structures of the*

− **GO Terms:**

- GO:0003677: *DNA binding*
- GO:0003823: *antigen binding*
- GO:0003824: *catalytic activity*
- GO:0004519: *endonuclease activity*
- GO:0005102: *receptor binding*

Figure 2. Entry from the BioLit PDB clone that has citations inthe BioLit database.

structure such that each letter in any given term links to the letter succeeding it. The full-text of the XML article is then read beginning with the abstract and ending with the end of the article body. The parser jumps to the beginning of every word in the article text, reading character by character. This structure is advantageous as it allows a search for all words in a single read of the document and it allows the reading to stop as soon as a word fails to have a character in the tree. The longest parse time that we have observed of parsing an article through a large ontology, such as the Gene Ontology, are about 30 s, although most articles are processed in under 1 s.

HARDWARE AND SOFTWARE DESIGN

The BioLit web server is staged from a single Dell PowerEdge SC145, which is hosted at the Skaggs School of Pharmacy and Pharmaceutical Sciences at UCSD. The machine is a dual-Opteron 2216 (Dual Core at 2.4 GHz) with 8 GB of RAM and 750 GB of RAIDed storage, which runs RedHat Enterprise Linux 5. At the present time, this machine is responsible for data mining, markup, front-end functionality, as well as serving a modified clone of the PDB web application.

In an effort to maximize performance and scalability of the software, the three main units of the project operate as discrete applications. The data mining portion is handled by a suite of in-house parsing scripts. The BioLit markup process is performed by a separate script. The front-end searching and visualization functionality is delivered by a J2EE web application running on Apache Tomcat 6. Much of the web-based, event-driven functionality is structured upon the YUI API. The dynamic data storage and sharing is handled by MySQL, while large static data are locally archived.

FUTURE DIRECTIONS

The value of the BioLit web server resides in the amount of metadata it can provide and in the ease in which those metadata can be retrieved. To increase the value of BioLit, we plan to parse the articles through all ontologies in the Open Biomedical Ontologies Foundry (20) and allow the user to specify the inclusion or exclusion of metadata by ontology. We also plan to expand our search for database identifiers to other major biological databases. In order to make these data more accessible to biocurators and text miners, we are implementing web services-based protocols that allow fetching of articles or metadata by source (database, ontology) and by term or ID.

We recognize that we are not experts in natural language processing and that the markup we provide is not truly semantic in nature (e.g. we do not infer relationships between terms and identifiers). We are open to collaboration in order to include this information in future versions.

CONCLUSIONS

We hope that the BioLit web server will establish an informative, yet transparent, connection between the data and the article describing the data and that effective use of this resource will provide new perspectives on both traditional literature and biological databases. We expect that literature will simply become another interface to biological data in a database and the database will recall appropriate literature—not in abstract or complete paper size chunks, but knowledge objects that annotate the data being examined. As authors become more aware of the possibilities offered by open access literature and by tools that can be used to standardize and highlight semantic content, we hope this awareness will manifest itself in the writing process and contribute to the accessibility of literature-based knowledge.

ACKNOWLEDGEMENTS

Special thanks go to Tim Beck, Bojan Beran, D. Jeff Milton, and Joe Mulvaney for assistance with system administration and installation of the PDB clone. Project funding was provided by National Science Foundation Award # 0544575. Funding to pay the Open Access publication charges for this article was provided by National Science Foundation Award #0544575.

Conflict of interest statement. None declared.

REFERENCES

1. Brown,P.O., Eisen,M.B. and Varmus,H.E. (2003) Why PLoS became a publisher. *PLoS Biol.*, **1**, E36.
2. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, **36**, D13–D21.
3. Douglas,S.M., Montelione,G.T. and Gerstein,M. (2005) PubNet: a flexible system for visualizing literature derived networks. *Genome Biol.*, **6**, R80.
4. Fink,L. and Bourne,P. (2007) Reinventing scholarly communication for the electronic age. *CTWatch Quarterly*, **3**, 26–31.
5. Bourne,P. (2005) Will a biological database be different from a biological journal? *PLoS Comput. Biol.*, **1**, 179–181.
6. MacCallum,C.J. (2007) When is open access not open access? *PLoS Biol.*, **5**, 2095–2097.
7. Seringhaus,M.R. and Gerstein,M.B. (2007) Publishing perishing? Towards tomorrow's information architecture. *BMC Bioinform.*, **8**, 17.
8. Eysenbach,G. (2006) The open access advantage. *J. Med. Int. Res.*, **8**, e8.
9. Eysenbach,G. (2006) Citation advantage of open access articles. *PLoS Biol.*, **4**, e157.
10. Watson,R. (2007) EC to promote open access publishing. *Br. Med. J.*, **334**, 389.
11. Kouranov,A., Xie,L., de la Cruz,J., Chen,L., Westbrook,J., Bourne,P.E. and Berman,H.M. (2006) The RCSB PDB information portal for structural genomics. *Nucleic Acids Res.*, **34**, D302–D305.
12. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
13. Hirschman,L., Yeh,A., Blaschke,C. and Valencia,A. (2005) Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinform.*, **6 (Suppl. 1)**, S1.
14. Doms,A. and Schroeder,M. (2005) GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res.*, **33**, W783–W786.
15. Vanteru,B.C., Shaik,J.S. and Yeasin,M. (2008) Semantically linking and browsing PubMed abstracts with gene ontology. *BMC Genomics*, **9 (Suppl. 1)**, S10.
16. Chen,D., Muller,H.M. and Sternberg,P.W. (2006) Automatic document classification of biological literature. *BMC Bioinform.*, **7**, 370.
17. Muller,H.M., Kenny,E.E. and Sternberg,P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.*, **2**, e309.
18. Rubin,D.L., Lewis,S.E., Mungall,C.J., Misra,S., Westerfield,M., Ashburner,M., Sim,I., Chute,C.G., Solbrig,H., Storey,M.A. *et al.* (2006) National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. *OMICS*, **10**, 185–198.
19. Yesilada,Y., Bechhofer,S. and Horan,B. (2007) COHSE: dynamic linking of web resources. Sun Microsystems TR-2007-167.
20. Smith,B., Ashburner,M., Rosse,C., Bard,J., Bug,W., Ceusters,W., Goldberg,L.J., Eilbeck,K., Ireland,A., Mungall,C.J. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.